氏　　名	大野　学
授与した学位	博　士
専攻分野の名称	理　学
学位授与番号	博甲第３４９１号
学位授与の日付	平成１９年　９月３０日
学位授与の要件	自然科学研究科資源管理科学専攻 （学位規則第４条第１項該当）
学位論文の題目	Studies on Variable Selection Methods Using Apriori Algorithm on Binary Data （２値データにおけるアプリオリアルゴリズムを利用した変数選択手法の研究）
論文審査委員	教授　垂水　共之　　教授　栗原　考次　　教授　梶原　毅

学位論文内容の要旨

Web 技術の発達により，インターネットを介して大量のデータを低コストで収集・蓄積することが可能になった．特に従来では困難であった，人を対象とした商品の選好や嗜好などのアンケート調査が低コストで実施出来るようになった．

そのようなアンケートデータには多くのカテゴリカルデータが含まれている．特に，商品の選好を問う質問項目では，好き／嫌いのような２値応答の項目が多く含まれる．また，アンケートデータに限らず，ＩＤつきＰＯＳデータのような顧客の購買履歴データでは，商品の買った／買わなかったという２値データが大量に存在する．ＰＯＳデータは，アンケートデータに比べてサイズが大きく、また変数の個数も膨大である．

このような背景があり，従来の統計学で想定されてきたデータのサイズを上回ってきており，従来の統計学の手法をそのまま適用させたときに，不都合が生じる場合ある．

カテゴリカルデータの分析のステップとして，まず初めに分割表を構築してデータの同時頻度を見ることが多い．しかしながら，従来のように変数が少ない場合は分割表の変数の組合せの個数が問題になることはなかったが，変数が多い場合は組合せ爆発の問題が生じる．

分割表において、変数同士が独立か関連があるかに注目した場合、ピアソンの独立性の検定を用いてその関係を判断することが多い．Brin *et al.* (1997) は 2 値データの分割表において，そのカイ２乗検定統計量が逆単調性をもつと述べた．これが正しければ，その性質を利用して効率的に独立な変数を見つけることが可能である．

2 値データにおいて単純に独立な変数の組合せを効率的に探す方法は，今まで提案されてない．その問題に対して逆単調性の特性を使うことは困難であると考える．しかし，逆単調性は効率的に変数を選択する上で重要な性質である．そこで，我々の提案する方法は変数の頑強性を考慮した変数選択法である．本論文で言う頑強性とは、オリジナルデータから選択された変数がそのデータに摂動を与えたデータからにも同じように選択される性質のことである．本論文は，独立に関する変数の頑強性が逆単調性の性質を持つことを数値実験によって確認した．逆単調性の性質が成り立てば，アプリオリアルゴリズム(Agrawal *et al.* 1994)によって効率的に変数を選択することができる．提案手法では，逆単調性をもつ変数の選択ステップにおいて，アプリオリアルゴリズムを利用した．

頑強性をもつ独立な変数の選択方法を応用し，顕示変数と局所独立な変数の組合せを選択する方法を提案した．我々は，Sakamoto(1991)が提案した AIC による分割表のモデル選択法を利用して，それらの変数を識別・選択した．最後に数値実験を用いて提案した方法の有効性を確認・検討した．

論文審査結果の要旨

コンビニ等のPOSデータに代表されるような大規模なデータは、これまでの統計学の想定を超えたデータ量となっている。このような膨大なデータから関連のある変数群を探索する手法は「データマイニング」と呼ばれる。データマイニングでは、関連のある「最適」な変数群を見つけることより、最適な変数群を含む変数群を効率よく探すことが求められている。

本研究では各個別商品の購入・未購入のように2値で表現されているデータから関連のある変数群を探す問題を取り扱っている。関連のある変数群の選択としては、変数の個数に対するある種の「単調性」がある場合に、アプリオリアルゴリズムと呼ばれるアルゴリズムを使うことにより、不要な探索を早めに打ち切り、効率よく探索することが可能となる。

本論文では、変数の独立性については、この単調性が成り立つことを示した。この拡張として条件付独立の場合にも単調性が成り立つことを示している。これより、ある変数で条件付けたときに独立となる「局所独立」な変数群の選択、および、独立な変数群にある変数を追加したとき、全体としては独立でなくなるような「顕示変数」の選択問題において、データに摂動を与えても、同じ目的の変数群が選択されるような「頑健」な手法を提案し、シミュレーションでその有効性を示した。

このように本論文では、2値大量データから局所独立な変数群の選択、ならびに顕示変数群の選択問題ついて、これまでにない新しいアルゴリズムを提案し、その有効性を示しており、計算機統計学の分野での価値は高い。

以上により本論文は博士に値すると判断した。